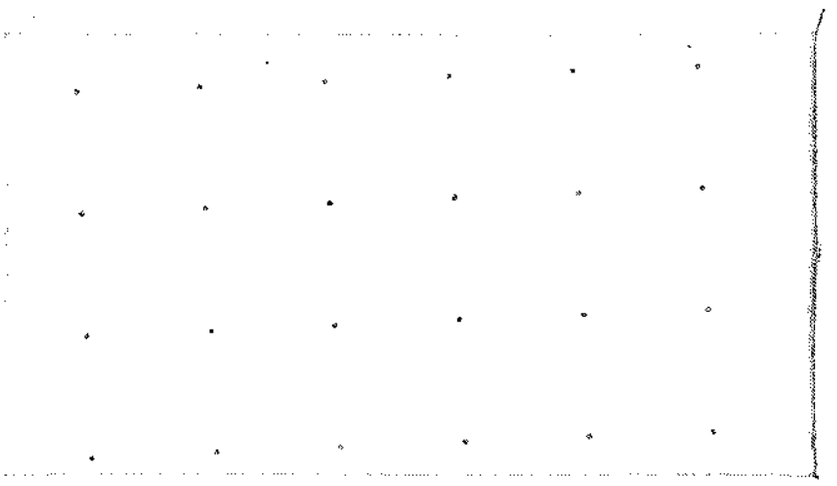
Log-linear models and goodness-of-fit statistics for categorical data in contingency table analysis

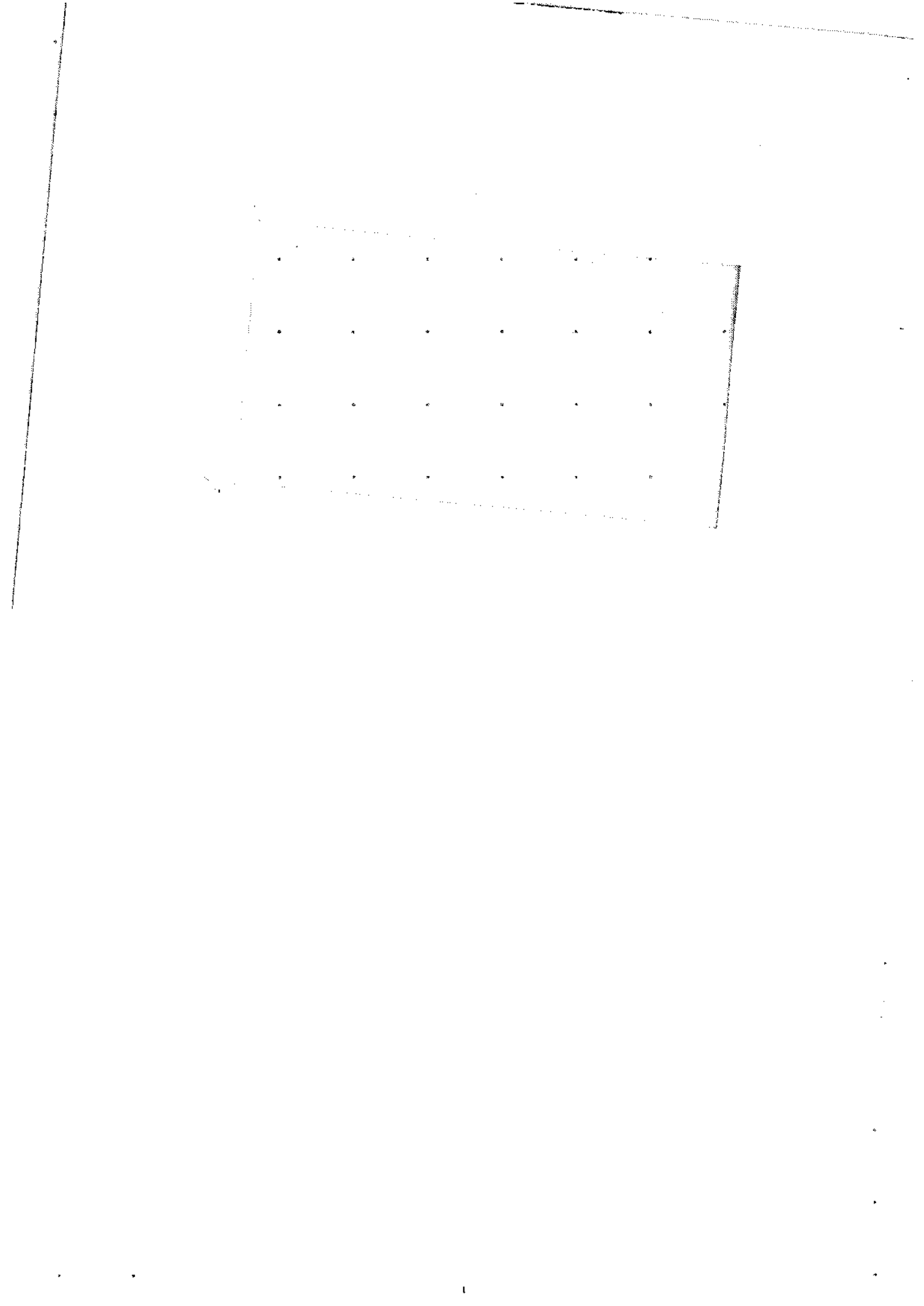
Brouwer, F.; Nijkamp, P.

1983

SERIE RESEARCHMEMORANDA



VRIJE UNIVERSITEIT
EKONOMISCHE FAKULTEIT
AMSTERDAM



Free University
A m s t e r d a m
Dept. of Economics
Secretary 020-5484933

LOG-LINEAR MODELS AND GOODNESS-OF-FIT STATISTICS FOR CATEGORICAL DATA IN CONTINGENCY TABLE ANALYSIS

An Application to Dutch Migration Data

F. Brouwer
P. Nijkamp

Researchmemorandum 1983-3 January '83

LOG-LINEAR MODELS AND GOODNESS-OF-FIT STATISTICS FOR CATEGORICAL DATA IN CONTINGENCY TABLE ANALYSIS

An Application to Dutch Migration Data

Floor Brouwer [1)]
Peter Nijkamp [2)]

[1)] Free University
Institute for Environmental Studies
P.O. Box 7161
1007 MC Amsterdam

[2)] Free University
Department of Economics
P.O. Box 7161
1007 MC Amsterdam.

This study was supported by the Netherlands Organization for the Advancement of Pure Research (ZWO), project-number 46-079.

Summary

Categorical data may be classified into a dichotomous and polychotomous data. In the present paper the use of such qualitative data will be discussed in the context of contingency table analysis.
A log-linear model and underlying statistical hypotheses are related to each other, especially in regard to nested or hierarchical models. The goodness-of-fit of different hierarchical models is judged by means of some asymptotic chi-squared test-statistics, viz. Pearsons test-statistic and the likelihood ratio-test.
A new issue in this paper is the discussion of the problem of a priori or structural zeros - in addition to sampling zeros - in log-linear analysis. This paper proposes a solution to deal with this problem.
The asnalysis is illustrated by means of an empirical application to inter-regional migration data in the Netherlands.

Contents:

Summary

1. Introduction 1

2. Contingency Table Analysis and Statistical Tests 2

3. Selection of a Hierarchical Model 8

4. Estimation of a Hierarchical Model 11

5. Sampling Zeros and Structural Zeros 14

6. Application to Migration Analysis 16

7. Conclusion 20

References 21

1. INTRODUCTION

Surveys often contain categorical data that have been classified in qualitative terms such as man or woman; married or single persons. These are examples of dichotomous classifications, which are measured on a nominal scale. This means that the variables or phenomena concerned can only be distinguished by their name or attribute. In order to classify such variables, numbers as well as other symbols can be used. Polychotomous or multiple classifications are also common, for example, a classification of people according to their income into low, middle or high income groups. The latter type of classified variables which can be ranked from low to high, is called ordinal data. Such ordinal data imply that it is known whether some observation has a higher or lower value than another one.

These categorical data will be discussed here in the context of contingency tables, while next log-linear models with their underlying statistical test-hypotheses will be developed to deal with this type of information. In this way, a contingency table can be interpreted as a sample from a multinomial distribution. The usual standard regression analysis and the contingency table analysis are both statistical models that are used for different kinds of data. Linear or non-linear regression models and many other econometric estimation procedures can be typified as analyzing a quantitative dependent variable as a linear or non-linear function of other independent variables. A main difference between regression procedures and log-linear models is that the latter type of modeling does not explicitly distinguish between response and explanatory variables: each variable can be analyzed simultaneously as a function of all others.

In this paper, a log-linear model and an underlying statistical hypothesis will be related to each other. Special attention will be paid to so called hierarchical models. The goodness-of-fit of the corresponding maximum-likelihood estimates of cell-elements will be tested by means of asymptotic chi-squared test-statistics, viz. Pearson's test-statistic and the likelihood-ratio statistic . The additivity property of the last one will also be demonstrated, which can be interpreted as a conditional hypothesis of adding new terms to the log-linear model. In addition, the problem of structural zeros in contingency tables will be formulated. A meaningful solution to deal with this problem appears to be the determination of restricted maximum-likelihood estimates. Some applications and results for regional migration data will be presented as well.

2. Contingency Table Analysis and Statistical Tests

The simplest form of a contingency is a 2x2 table, in which the entries reflect the observations or frequencies of occurrence of each pair of categories. A survey with sample size N can be represented in the following way.

	1	2	total
1	n_{11}	n_{12}	$n_{1\cdot}$
2	n_{21}	n_{22}	$n_{2\cdot}$
total	$n_{\cdot 1}$	$n_{\cdot 2}$	N

Table 1. A 2x2 Contingency Table

This 2x2 two-way table can be generalized to an I x J table with I rows and J columns, and cell-elements n_{ij}, i=1,...,I; j=1,...,J. The row and column sums, called the marginal totals, are respectively equal to:

$$n_{i\cdot} = \sum_{j=1}^{J} n_{ij} \quad , \; i=1,\ldots,I \tag{1}$$

$$n_{\cdot j} = \sum_{i=1}^{I} n_{ij} \quad , \; j=1,\ldots,J \tag{2}$$

Clearly, the sum of all these elements is:

$$N = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} = \sum_{i=1}^{I} n_{i\cdot} = \sum_{j=1}^{J} n_{\cdot j} \tag{3}$$

This two-way table can easily be generalized to a higher order one, i.e., a three-way table with three components and elements n_{ijk}, i=1,...,I; j=1,...,J; k=1,...,K.

Applications of contingency tables analyses in geography, demography, biology, medicine and sociology can be found in among others Abrahamse and Lammerts van Bueren, 1979, 1980a, 1980b; Aitkin, 1979, 1980; Cox, 1970; Everitt, 1977; Fingleton, 1981; McCullagh, 1980; Pannekoek, 1980; Willekens, 1980, 1982.

The analysis of contingency tables usually starts with assumptions about the characteristics of the relationship between the observed variables, which will be tested thereafter.

When all cell elements in the contingency table are assumed to be independent, the probability of occurrence of some event can be expressed in terms of the corresponding marginal totals. This means there are no interaction effects between pairs of variables. For example, in a three-way contingency table, a test on complete independence in terms of probabilities of occurrence is given by:

$$H_o: P^*_{ijk} = P_{i..} P_{.j.} P_{..k} \quad \text{for all } i,j \text{ and } k \text{ with } P_{i..}, P_{.j.}, P_{..k} > 0 \tag{4}$$

$$H_1: P^*_{ijk} \neq P_{i..} P_{.j.} P_{..k} \quad \text{for at least one pair of combinations } (i,j,k) \text{ where } i=1,\ldots,I;\ j=1,\ldots,J;\ k=1,\ldots,K$$

When the null hypothesis will not be rejected, the probabilities of events are represented in terms of products of the corresponding marginal probabilities. This hypothesis can simply be transformed in terms of expected frequency estimates n^*_{ijk}:

$$H_o: n^*_{ijk} = N \frac{n_{i..}}{N} \frac{n_{.j.}}{N} \frac{n_{..k}}{N} = \frac{n_{i..} n_{.j.} n_{..k}}{N^2} \quad \text{for all } i,j \text{ and } k \tag{5}$$

$$H_1: n^*_{ijk} \neq \frac{n_{i..} n_{.j.} n_{..k}}{N^2} \quad \text{for at least one pair } (i,j,k)$$

The elements n^*_{ijk} are maximum - likelihood estimates, given the conditions of the null-hypothesis of mutual independece. The marginal category totals are assumed to be given. When logarithms from (5) are taken - again given the condition of the null-hypothesis of mutual independence -, one gets:

$$\ln n^*_{ijk} = \ln n_{i..} + \ln n_{.j.} + \ln n_{..k} - 2 \ln N; \quad i=1,..,I; j=1,..J; k=1,..,K. \tag{6}$$

This can be rewritten in terms of an overall mean and main effects. Because independency is assumed in (5), there are no first-order or interaction effects between the variables. Birch (1963) has shown that (6) can be formulated as a log-linear model:

$$\ln n^{*}_{ijk} = \mu + \mu_1(i) + \mu_2(j) + \mu_3(k) \quad \text{for all } i,j \text{ and } k \tag{7}$$

where:

$$\begin{aligned}
\mu &= \frac{1}{IJK} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \ln n^{*}_{ijk} \\
\mu_1(i) &= \frac{1}{JK} \sum_{j=1}^{J} \sum_{k=1}^{K} \ln n^{*}_{ijk} - \mu \quad , i=1,\ldots,I \\
\mu_2(j) &= \frac{1}{IK} \sum_{i=1}^{I} \sum_{k=1}^{K} \ln n^{*}_{ijk} - \mu \quad , j=1,\ldots,J \\
\mu_3(k) &= \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \ln n^{*}_{ijk} - \mu \quad , k=1,\ldots,K
\end{aligned} \tag{8}$$

μ is the overall mean, where $\mu_1(i)$, $\mu_2(j)$, and $\mu_3(k)$ give the main effects from component 1,2 and 3 respectively.

One main advantage of taking logarithms from multiplicative statistical hypotheses like (5) is the emergence of a log-linear model with additive components. Given this, the effect of one single variable on the cell frequencies can be measured. The equivalence of formulae (6) and (7) can easily be proved by means of the conditional hypothesis of mutual independency as represented in (5).

In the same way other types of log-linear models with their underlying statistical hypotheses can be developed (see also Brouwer, 1982; Payne, 1977). Suppose again a IxJxK contingency table with sample size N where the following null-hypotheses with two-sided alternative hypotheses will be tested successively:

$$H_o : n^*_{ijk} = \frac{N}{IJK} \qquad \text{for all i,j and k} \qquad (9a)$$

$$H_o : n^*_{ijk} = \frac{n_{i..}\, n_{.j.}\, n_{..k}}{N^2} \qquad \text{for all i,j and k} \qquad (9b)$$

$$H_o : n^*_{ijk} = \frac{n_{ij.}\, n_{..k}}{N} \qquad \text{for all i,j and k} \qquad (9c)$$

$$H_0 : n^*_{ijk} = \frac{n_{ij.}\, n_{i.k}}{n_{i..}} \qquad \text{for all i,j and k} \qquad (9d)$$

$$H_o : \begin{cases} n^*_{ij.} = n_{ij.} \\ n^*_{.jk} = n_{.jk} \\ n^*_{i.k} = n_{i.k} \end{cases} \qquad \text{for all i,j and k} \qquad (9e)$$

The first test assumes that the cell frequencies are equally distributed through the IJK elements. Test (9b) assumes that there are no interaction effects between the variables, i.e. all variables are mutual independent from each other.Multiple independence has been assumed in formula (9c): the joint variable one and two is independent of the third variable. Test (9d) is related to conditional independence: the second and third variable are independent, given the level of the first variable. The last test supposes that there is no second-order interaction effect between the variables.
These five hypotheses imply the following log-linear models successively:

$$\ln n^*_{ijk} = \mu \qquad (10a)$$

$$\ln n^*_{ijk} = \mu + \mu_1(i) + \mu_2(j) + \mu_3(k) \qquad (10b)$$

$$\ln n^*_{ijk} = \mu + \mu_1(i) + \mu_2(j) + \mu_3(k) + \mu_{12}(ij) \qquad (10c)$$

$$\ln n^*_{ijk} = \mu + \mu_1(i) + \mu_2(j) + \mu_3(k) + \mu_{12}(ij) + \mu_{13}(ik) \qquad (10d)$$

$$\ln n^*_{ijk} = \mu + \mu_1(i) + \mu_2(j) + \mu_3(k) + \mu_{12}(ij) + \mu_{13}(ik) + \mu_{23}(jk) \qquad (10e)$$

These five log-linear models hold - given the statistical hypotheses from formulae (9a)-(9e) - for all values of i, j, k with i=1, ..., I ; j=1,...,J;k=1,...,K. They will be called hierarchical log-linear models, because the inclusion of a given order effect also assumes the inclusion of all lower-order effects. The model that includes all possible μ terms is called the saturated model; for the three-way contingency table it has the following shape:

$$\ln n^{*}_{ijk} = \mu + \mu_{1}(i) + \mu_{2}(j) + \mu_{3}(k) + \mu_{12}(ij) + \mu_{13}(ik) + \mu_{23}(jk) + \mu_{123}(ijk) \qquad (11)$$

This model is the most general model that can be fitted to a three-way contingency table; it follows from the hypothesis that the expected value of cell-frequencies is identical to the observed values, or:

$$H_{o} : n^{*}_{ijk} = n_{ijk} \quad \text{for all } i,j \text{ and } k \qquad (12)$$

For the saturated model, the number of cell-elements is equal to the number of μ-variables. Main effects and first-order effects in formulae (10a-e) are determined by (8) and a typical first-order interaction effect is:

$$\mu_{12}(ij) = \frac{1}{K}\sum_{k=1}^{K} \ln n^{*}_{ijk} - \frac{1}{JK}\sum_{j=1}^{J}\sum_{k=1}^{K} \ln n^{*}_{ijk} - \frac{1}{IK}\sum_{i=1}^{I}\sum_{k=1}^{K} \ln n^{*}_{ijk} + \mu \qquad (13)$$

Since the main and higher-order effects are determined by deviations from a mean, the μ-terms have the following constraints:

$$\begin{aligned} &\sum_{i} \mu_{1}(i) = \sum_{j} \mu_{2}(j) = \sum_{k} \mu_{3}(k) = 0 \\ &\sum_{i} \mu_{12}(ij) = \sum_{j} \mu_{12}(ij) = 0 \\ &\sum_{i} \mu_{123}(ijk) = \sum_{j} \mu_{123}(ijk) = \sum_{k} \mu_{123}(ijk) = 0 \end{aligned} \qquad \begin{aligned} &i=1,\ldots,I \\ &j=1,\ldots,J \\ &k=1,\ldots,K \end{aligned} \qquad (14)$$

The same constraints hold for other μ-terms.

Given the condition that the null-hypothesis will not be rejected, the expected values of cell-frequencies in formula (9) are maximum likelihood estimates. It has been shown (see also Birch, 1963) that the marginal totals are sufficient statistics. In a hierarchical log-linear model in a three-way table where, for example second-order interaction effects are assumed to be zero (i.e., $\mu_{123} = 0$), like in (10e), the sufficient statistics are N, $n_{i..}$, $n_{j.}$, $n_{..k}$, $n_{ij.}$, $n_{i.k}$, $n_{.jk}$.
Because of a hierarchical model the minimum set of sufficient statistics is $n_{ij.}$, $n_{i.k}$ and $n_{.jk}$. A very simple and useful iterative
method of obtaining maximum likelihood estimates of the expected cell frequencies is an iterative scaling procedure (see also Fienberg, 1970; Payne, 1977). This procedure employs the minimum set of sufficient statistics. When a statistical hypothesis like (10e) is assumed, the maximum likelihood estimates n^*_{ijk} must satisfy three constraints. These constraints follow immediately from the sufficient statistics and are:

$$\begin{aligned} n^*_{ij.} &= n_{ij.} \\ n^*_{i.k} &= n_{i.k} \qquad \text{for all i, j and k} \\ n^*_{.jk} &= n_{.jk} \end{aligned} \tag{15}$$

The iterative procedure starts with initial values

$$n^{(0)}_{ijk} = 1$$

for each cell-element and these constraints lead to the following iterative procedure in stage s.

$$\text{step 1.} \quad n^{(3s+1)}_{ijk} = n^{(3s)}_{ijk} \cdot \frac{n_{ij.}}{n^{(3s)}_{ij.}} \qquad \text{for all i,j and k}$$

$$\text{step 2.} \quad n^{(3s+2)}_{ijk} = n^{(3s+1)}_{ijk} \cdot \frac{n_{i.k}}{n^{(3s+1)}_{i.k}} \qquad \text{for all i,j and k} \tag{16}$$

$$\text{step 3.} \quad n^{(3s+3)}_{ijk} = n^{(3s+2)}_{ijk} \cdot \frac{n_{.jk}}{n^{(3s+2)}_{.jk}} \qquad \text{for all i,j and k}$$

These three steps represent the s-th cycle of the iteration. When the difference between successive estimates of n_{ijk} for each cell element is less than a preliminar chosen small value, say 0.01, the iteration procedure is terminated and convergent maximum likelihood solutions are obtained. When the termination criterium does not hold, then s+1:=s. In that case one has to start again with step 1 until the convergent criterium holds for all elements n_{ijk}.

3. Selection of a Hierarchical Model

In the previous section, some log-linear models and their statistical hypotheses have been discussed. In the present section, the fit of these models will be mutually compared in order to arrive at the selection of a specific model.

For each type of contingency tables two extreme models may be considered:

- a minimal model with the smallest set of parameters, i.e., the overall mean effect. See for example model (10a) for a three-way table.
- a saturated model in which all possible parameters are included and the expected values of cell frequencies are by definition identical to the observed ones. See for example model (11) for a three-way table.

The fit of the expected cell frequencies with the observed frequencies under some statistical hypotheses uses asymptotically chi-squared distributed test-statistics. On the one hand we have the test-statistic, developed by Pearson in 1904, where the sum of squared deviations between observed and expected frequencies are related to the expected frequencies, i.e.

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{(n_{ijk} - n^*_{ijk})^2}{n^*_{ijk}} \qquad (17)$$

In addition, the likelihood ratio test is developed in the same way (see also Mood et al.,1974) and its test-statistic is determined by the likelihood ratio λ, with:

$$\lambda = \frac{L_0(p^*)}{L(p^*)} \qquad (18)$$

$L(p^*)$ is the unrestricted maximum value of the likelihood function and $Lo(p^*)$ is the maximum of the likelihood function conditional to some given null hypothesis. The likelihood ratio-test $-2 \ln \lambda$ which can also be approximated a-symptotically by a chi-squared distribution, is given by:

$$G^2 = -2 \ln \lambda = 2 \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} n_{ijk} \ln \left(\frac{n_{ijk}}{n^*_{ijk}} \right) \qquad (19)$$

The number of degrees of freedom is defined by the total number of independent terms (see also Everitt, 1977; Kullback, 1968)[1]; in the case of mutual independence it is equal to (I-1)(J-1)(K-1) because of the given marginal totals. By means of (17) and (19), we are able to test a certain null hypothesis. If the χ^2-value is larger than some critical value, the null hypothesis will be rejected. When for example the null hypothesis given in (9b) is rejected by means of the test statistics (17) or (19), we can conclude that there is no mutual independence between the components. In that case, the interaction measures between categories have to be computed which will give smaller values of both Pearson's chi-squared statistic and the likelihood ratio-test.

These two test-statistics are not equal to each other, because they have an asymptotic chi-squared distribution. The likelihood ratio test is preferred to the Pearson's chi-square statistic because of its additive character (see also Brouwer, 1982; Goodman, 1972; Payne, 1977). In that way, a statistical test can be developed to test whether the addition of one or more parameters in a hierarchical model gives an improvement of the model. Suppose for example the following log-linear model:

$$\ln n^{*}_{ijk} = \mu + \mu_1 (i) + \mu_2 (j) + \mu_3 (k) \tag{20a}$$

$$\ln n^{**}_{ijk} = \mu + \mu_1 (i) + \mu_2 (j) + \mu_3 (k) + \mu_{12} (ij) \tag{20b}$$

The difference between the respective likelihood ratio test-statistics is also distributed as χ^2 with one degree of freedom and provides a conditional test of the hypothesis μ_{12} (ij) = 0. The additive character of the likelihood ratio test follows by:

$$-2 \ln \lambda = 2 \sum_{i,j,k} n_{ijk} \ln \left(\frac{n_{ijk}}{n^{*}_{ijk}} \right) = 2 \sum_{i,j,k} n_{ijk} \ln \left(\frac{n_{ijk}}{n^{**}_{ijk}} \; \frac{n^{**}_{ijk}}{n^{*}_{ijk}} \right)$$

$$= 2 \sum_{i,j,k} n_{ijk} \ln \left(\frac{n_{ijk}}{n^{**}_{ijk}} \right) + 2 \sum_{i,j,k} n_{ijk} \ln \left(\frac{n^{**}_{ijk}}{n^{*}_{ijk}} \right) \tag{21}$$

[1]) Note: the number of independent terms is equal to the difference between the number of cell elements and the number of parameters to be fitted.

The expected cell frequencies n_{ijk} are assumed to satisfy a log-linear model which is a special case of the hierarchical model with expected cell frequencies n^{**}_{ijk}. Then the last term in (21) can be used to test whether the differences between the expected values n^{*}_{ijk} and n^{**}_{ijk} are simply due to random variation, given that the true expected frequencies are n^{**}_{ijk}. Its number of degrees of freedom are equal to the difference in degrees of freedom between the two models. From this follows that the likelihood ratio statistic from model (10a), called $G^2(a)$, can be decomposed into the following additive terms:

$$G^2(a) = \left(G^2(a) - G^2(b)\right) + \left(G^2(b) - G^2(c)\right) + \left(G^2(c) - G^2(e)\right) + G^2(e) \qquad (22)$$

Each pair of components in the brackets has an aymptotic χ^2-distribution with appropriate degrees of freedom.

The fit of a model can be represented by the G^2-test value. A large G^2 value can be interpreted as a bad fit. Goodman (1972) interprets this likelihood ratio statistic as the variation in the log frequencies which are unexplained by the model. To have a measure that represents the relative improvement of a log-linear model, he determined:

$$\frac{G^2(i) - G^2(k)}{G^2(i)} \qquad (23)$$

This can be interpreted as the relative improvement of some hierarchical model related to a model with less parameters. When a certain model i is taken as the minimal model, coefficient (23) corresponds to the multiple correlation coefficient (R^2 in regression analysis). The interval $\lfloor 0,1 \rfloor$ is the range of possible values of this coefficient. It attains its maximum value when some model has an exact fit with the observed data and it has a value 0 when the model will not improve the goodness of fit related to the original model.

When model (10b) is compared with the saturated one from (11), the sum of first- and second-order interaction-effects, i.e, μ_{12} (ij) + μ_{13} (ik) + μ_{23} (jk) + μ_{23} (ijk), is called the disturbance term in (10b). In many practical applications there are only a limited number of cell-elements, and the number of components is reduced by means of some statistical hypothesis. But when there are a lot of possible first-, second- and higher-order effects, it is necessary to reduce the number of components in a log-linear model with the sum of the other effects (called the disturbance term). There is an erratic movement of the disturbance term when it consists of a finite number of components. But when the number of components of the disturbance term become infinite, it can be adjusted by a known distribution function like the normal distribution. This normal distribution fucntion complies with the central limit theorem.
In this way the disturbance term can be specified exactly with this type of analysis, where in regression analysis the distrubance term consists of the unobserved and unknown variables.

4.Estimation of a Hierarchical Model

In the previous section unique maximum likelihood estimates for the expected values of cell-frequencies have been developed. These maximum likelihood estimates exist for this model, given a statistical independency of the observations. These may be generated by one of the following sampling procedures (see also Bishop, et al., 1975; Fienberg, 1977):

- independent Poisson sampling
- simple multinomial sampling
- product multinomial sampling

The μ-terms, which represent main effects and higher-order effects, will be estimated in this section. We are interested, for example, in the difference μ_1 (1) - μ_1 (2), when the first-order effect is at level 1 and 2, respectively. From this, it can be tested whether the effect at level 1 on the log frequencies is significantly different from the effect of being at level 2.
The model for a three-way I x J x K multiway-table with P parameters to be estimated is (see also Payne, 1977):

$$y = X\mu + e \qquad (24)$$

The estimates of μ result from an ordinary least-squares estimation procedure:

$$\mu^* = (X'X)^{-1} X'y \qquad (25)$$

where

y is a (I x J x K) x 1 vector of log estimated frequencies
μ is a P x 1 vector of parameters to be estimated
X is a (I x J x K) x P design matrix with elements determined by the parameters required.
e is a (I x J x K) x 1 vector of error terms.

Because of the restrictions on the parameters [see also formula (14)], the model can also be rewritten in terms of redefined parameters, called $\beta_1 \ldots, \beta_p$. Suppose a 2x2x2 multiway-table. Then the reformulated model becomes:

$$\begin{bmatrix} \ln n^*_{111} \\ \ln n^*_{112} \\ \ln n^*_{121} \\ \ln n^*_{122} \\ \ln n^*_{211} \\ \ln n^*_{212} \\ \ln n^*_{221} \\ \ln n^*_{222} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & -1 & 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \\ \beta_7 \\ \beta_8 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \\ e_8 \end{bmatrix} \qquad (26)$$

The β terms can uniquely be rewritten in terms of μ elements which follow simply by substition (see also Payne, 1977). β_1 corresponds to the mean-effect μ, while β_2, β_3 and β_4 correspond to differences $\mu_1(1) - \mu_1(2)$, $\mu_2(1) - \mu_2(2)$, $\mu_3(1) - \mu_3(2)$, respectively. Note that the final column of the design matrix for β_8, which represents the second-order interaction-effect, is simply the product of the columns for the corresponding main effects and

first-order interaction effects. This corresponds to the interpretation of β_8 as the difference between two second-order interaction effects, viz., μ_{123} (111) - μ_{123} (112).

With the linear relationships in (26), the β terms can be rewritten in terms of the log expected frequencies, i.e.,

$$\beta_1 = \frac{1}{8}\left[\ln n^*_{111} + \ln n^*_{112} + \ln n^*_{121} + \ln n^*_{122} + \ln n^*_{211} + \ln n^*_{212} + \ln n^*_{221} + \ln n^*_{222}\right]$$

$$\beta_2 = \frac{1}{8}\left[\ln n^*_{111} + \ln n^*_{112} + \ln n^*_{121} + \ln n^*_{122} - \ln n^*_{211} - \ln n^*_{212} - \ln n^*_{221} - \ln n^*_{222}\right]$$

$$\beta_3 = \frac{1}{8}\left[\ln n^*_{111} + \ln n^*_{112} - \ln n^*_{121} - \ln n^*_{122} + \ln n^*_{211} + \ln n^*_{212} - \ln n^*_{221} - \ln n^*_{222}\right]$$

$$\beta_4 = \frac{1}{8}\left[\ln n^*_{111} - \ln n^*_{112} + \ln n^*_{121} - \ln n^*_{122} + \ln n^*_{211} - \ln n^*_{212} + \ln n^*_{221} - \ln n^*_{222}\right] \qquad (27)$$

$$\beta_5 = \frac{1}{8}\left[\ln n^*_{111} + \ln n^*_{112} - \ln n^*_{121} - \ln n^*_{122} - \ln n^*_{211} - \ln n^*_{212} + \ln n^*_{221} + \ln n^*_{222}\right]$$

$$\beta_6 = \frac{1}{8}\left[\ln n^*_{111} - \ln n^*_{112} + \ln n^*_{121} - \ln n^*_{122} - \ln n^*_{211} + \ln n^*_{212} - \ln n^*_{221} + \ln n^*_{222}\right]$$

$$\beta_7 = \frac{1}{8}\left[\ln n^*_{111} - \ln n^*_{112} - \ln n^*_{121} + \ln n^*_{122} + \ln n^*_{211} - \ln n^*_{212} - \ln n^*_{221} + \ln n^*_{222}\right]$$

$$\beta_8 = \frac{1}{8}\left[\ln n^*_{111} - \ln n^*_{112} - \ln n^*_{121} + \ln n^*_{122} - \ln n^*_{211} + \ln n^*_{212} + \ln n^*_{221} - \ln n^*_{222}\right]$$

Goodman (1972) has given a proof that the variance of the parameter estimates (27) is:

$$s^2 = \sum_{i,j,k} \frac{1}{64} / \ln n^*_{ijk} \qquad (28)$$

These variances can be used to test the hypothesis that some $\beta_i=0$, $i=1,\ldots,8$.

5. Sampling Zeros and Structural Zeros

In this section the problem of zero cell-elements will be dealt with. Multi-way-tables where some cells are ignored are called incomplete tables.
There are two possibilities with zero elements in contingency tables: fixed and sampling zeros (see also Bishop et al., 1975; Fienberg, 1977; Scholten, 1982; Willekens, 1982).
Sampling zeros may occur because of a relative small cell probability. They may vanish,at least in a theoretical way, when the sample size will sufficiently be increased, and do not necessarily have to be equal to zero. In contrast to sampling zeros, fixed(or structural a priori) zeros occur if some cell-elements are impossible or excluded from the analysis, and if cell elements are known to have a priori a zero value.
While the sampling zeros may occur because of a relatively small sample size when compared with the large number of cell-elements, they may have non-zero expected values. Evidently it will become problematic if one cannot distinguish between sampling zeros and structural zeros in a model, as then all zeros are treated as if they were structural zeros. One way out will be to replace a sampling zero by a small number before the log-linear analysis will be applied. Cox (1970) gives a proof that the addition of 0.5 to zero elements will give satisfactory asymptotic results (i.e., only for a large sample size).
Another method is suggested by Fienberg and Holland (1970) where a value ε will be added to zero cell-elements. This value depends on the observations and is defined as:

$$\varepsilon = \frac{N^2 - \sum_{i,j,k} n^2_{ijk}}{IJK \sum_{i,j,k} n^2_{ijk} - N^2} \qquad (29)$$

with a IxJxK contingency table with elements n_{ijk} and sample size N. ε will be zero when some cell-element n_{ijk} is equal to N and all other cell-elements are zero (see also Pannekoek, 1980).

There is a number of standard computer packages which provide in various ways solutions for the problem of zero cell elements. A widely used programme like ECTA (Everymans Contingency Table Analysis) adds a constant value to all zero cell frequencies; the GLIM (Generalized Linear Interactive Modelling) handles the zero frequencies correctly, by means of a restricted maximum likelihood procedure but a real disadvantage of GLIM is its unability to deal with many dimensions in this programme.

In the following, the expected values of incomplete tables (with structural zeros) will be estimated. A good illustration of the occurrence of structural zeros is migration analysis where intraregional migration is ignored. In that case the diagonal of a table consists of zero elements and the null hypothesis of the absence of interactions outside the main diagonal can be tested. This will also be called a quasi-independence model; this is like dependency to be tested on the non-zero elements of a table.

An algorithm is used by means of which iterative solutions are adjusted in a biproportional way to the given set of marginal category totals. This algorithm is an iterative procedure called the Deming-Stephan iterative proportional fitting procedure (see also Bishop et al., 1975). In this procedure, the cell-elements are partioned into a set S containing non-zero elements and a set of elements with structural zeros. The algorithm consists of an initial condition and a cycle of iterations :

$$n_{ijk}^{(o)} = \begin{cases} 1 & \text{if } (i,j,k) \in S \\ 0 & \text{in other cases} \end{cases} \qquad \text{for all } i,j \text{ and } k \qquad (30)$$

(30) is the initial condition where S represents the non-zero cell-elements. The s-th cycle of the iteration procedure consists of three steps similar to the one represented in formula (16) above. Because of the initial condition, the expected value of cell-elements containing structural zeros also becomes zero, and restricted maximum likelihood estimates are obtained.

The degrees of freedom also have to be determined when the goodness-of-fit of this quasi-independence model will be tested. If the table consists of IJK elements and s structural zeros, only (I-1)(J-1)(K-1)- s independent elements (called the degrees of freedom) remain in the quasi-independence model.

6. Application to Migration Analysis

Table 1 represents interregional and intraregional migration flows, subdivided into male and female persons for the Dutch provinces Groningen and Friesland. The row-elements are incoming migrants and the column elements are out-going migrants.

	Male		Female	
	Groningen	Friesland	Groningen	Friesland
Groningen	10074	1328	8089	1099
Friesland	1260	10137	999	8607

Table 1 2x2x2 Contingency Table of Provincial Migration Flows in 1978

It is easlily seen that 10074 male persons out of the total sample survey of 41593 move within the province of Groningen. In addition, there are 22799 male migrants and 18794 female migrants.

Given the observed frequencies from table 1, the null-hypotheses will be tested for formulae (9a)-(9e). The expected cell-frequencies with goodness-of-fit statistics are given in a lexicographic order in table 2. The critical chi-square value has a significance level of 95 percent.

		Expected cell-frequencies				
(i,j,k)	observed	(9a)	(9b)	(9c)	(9d)	(9e)
(1,1,1)	10074	5199.125	5541.533	9955.960	10058.015	10083.143
(1,1,2)	8089	5199.125	4568.076	8207.040	8104.985	8079.857
(1,2,1)	1328	5199.125	5744.775	1238.260	1343.985	1319.062
(1,2,2)	1099	5199.125	4735.616	1020.740	1083.015	1107.938
(2,1,1)	1260	5199.125	5652.687	1330.348	1225.816	1250.857
(2,1,2)	999	5199.125	4659.704	1096.652	1033.506	1008.143
(2,2,1)	10137	5199.125	5860.005	10274.432	10171.184	10145.931
(2,2,2)	8607	5199.125	4830.604	8469.568	8575.494	8598.069
Pearson's X^2		25593.99	24972.68	32.08	2.82	0.31
Likelihood ratio test		32355.95	28380.33	32.12	3.18	0.33
Degrees of Freedom		7	4	3	2	1
Critical X^2-value		14.07	9.49	7.81	5.99	3.84

Table 2. Expected Cell-frequencies of Migration Flows

It can easily be proved that the expected cell-frequencies from the saturated model are identical to the observed frequencies.In that case all information from the observations will be used.Models (9a)-(9c) appear to have a poor fit,and model (9d) should be chosen.In model (9d), the interaction between destination and sex is also included.
The relative improvement of this log-linear model can be assesed by using the Goodman-index (see also formula (23)); the results are given in table 3.

Model	(9a)-(9b)	(9b)-(9c)	(9c)-(9d)	(9d)-(9e)
Goodman-index	0.1229	0.9989	0.9010	0.8962

(9a)-(9c)	(9a)-(9d)	(9a)-(9e)
0.9990	0.9999	1.0000

Table 3. Goodman-index of the Log-linear Models

Addition of new parameters always leads to lower chi-square values because information from the observations (like inter-relationships between the variables) will be used in the log-linear model. When model (9c) is related to model (9b), the goodness-of-fit statistics show that addition of the relationship between the first and second variable (region of origin and destination) improves the model substantially.

The β-parameters, represent the main effects and interaction effects successively and which can be determined by formula (27) (with n^*_{ijk} the expected cell-frequencies from the saturated model), then become:

β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
8.092	0.0099	-0.0270	0.1005	1.0326	0.0016	0.0123	0.2644

β_1 represents the mean effect; β_2, β_3 and β_4 the main effects; β_5, β_6 and β_7 the first order interaction effect and β_8 the second order interaction effect.

The standard deviation of these parameters is equal to 0.1253 which follows simply by substitution into formula (28). The significant parameters are β_1 (mean effect), β_5 (interaction effect between the region of origin and destination) and β_8 (second-order interaction-effect between sex and region of origin and destination).

In the next exercise, a two-way table will be tested where the region of destination and origin in a migration analysis are related to each other. This analysis will be carried out for an interregional migration table for the Netherlands as a whole. The regions are subdivided into the Northern, Eastern, Western, South-Western and Southern part of the Netherlands. The interregional and intraregional migration flows in 1978 are represented in table 4[1]).

	Origin					
Destination	North	East	West	S-West	South	total
North	50829	8953	12606	283	1913	74584
East	7820	61611	25242	801	9493	104967
West	8760	20890	211517	3100	15719	259986
S-West	269	826	5385	8094	2161	16735
South	1780	9876	20157	1745	87858	121416
total	69458	102156	274907	14023	117144	577688

Table 4. Interregional and Intraregional Migration Flows in 1978.

In an analogous manner, the following hypothesis of complete independence will be tested:

$$H_0 : \begin{cases} n^*_{i.} = n_{i.} \\ n^*_{.j} = n_{.j} \end{cases} \qquad \text{for all } i, j;\ i, j = 1,..,5 \qquad (31)$$

This iterative procedure gives expected values of cell-frequencies which are convergent after one iteration. These values are represented in table 5.

	Origin				
Destination	North	East	West	S-West	South
North	8968	13189	35492	1811	15124
East	12621	18562	49951	2548	21285
West	31259	45975	123721	6311	52720
S-West	2012	2959	7964	406	3394
South	14598	21471	57779	2947	24621

Table 5. Expected Values of Interregional and Intraregional Migration Flows.

[1]) Source: Statistics of internal migration in 1978, Central Bureau of Statistics, State Publishing Company, The Hague, 1981.

Pearson's chi-square statistic is equal to 644454.95 and the hypothesis of complete independence will be rejected for the 16 degrees of freedom and a 95 percent significance level of this statistic.

The relationships between the region of origin and destination can also be related to each other when the intraregional migration flows will be ignored. This will be done because of the different explanatory character of interregional and intraregional migration flows. In that way, the main diagonal of the migration table from table 4 contains zero elements. This is evidently a situation with structural zeros. The Deming-Stephan iterative proportional fitting procedure gives the expected values, given the hypothesis of quasi-independency represented in table 6. Convergent solutions are obtained after 10 runs.

	origin					
Destination	North	East	West	S-West	South	total
North	0	6344	12577	741	4092	23754
East	5354	0	27426	1560	9016	43356
West	8709	22453	0	2540	14767	48469
S-West	832	2133	4265	0	1411	8641
South	3734	9615	19122	1088	0	33559
Total	18629	40545	63390	5929	29286	157779

Table 6. Expected Values of Migration Flows Conditional to Quasi-independency.

Pearson's chi-square statistic and the likelihood ratio-test statistic are equal to 7844.38 and 7503.68, respectively with 16-5 = 11 degrees of freedom; the hypothesis of quasi-independence will be rejected at a 95 percent significance level.

It has been indicated in section 5 that some standard computer packages use structural zeros as if these were sample zero cell frequencies. Table 7 gives the expected values of migration flows conditional to complete independence when a small values ε - say 0,5 - is added to the zero cell frequencies.

Destination	origin North	East	West	S-West	South	total
North	2805	6105	9544	893	4409	23756
East	5119	11141	17419	1629	8048	43356
West	5723	12455	19473	1821	8997	48469
S-West	1020	2220	3472	325	1604	8641
South	3962	8624	13482	1261	6229	33558
Total	18629	40545	63390	5929	29287	157780

Table 7. Expected Values of Migration Flows Conditional to Complete Independency.

In addition the value of ε - no matter whether ε is equal to or smaller than 0.5 - has only minor effects on the expected values of the cell frequencies. When table 6 and table 7 are compared with each other, it becomes clear that the consequences of using structural zeros as if these were sampling zeros are rather significant.

7. Conclusion

This paper deals with contingency table analysis for categorical data, measured on a nominal or an ordinal scale. A statistical hypothesis about the relationship between the marginal category totals gives a hierarchical log-linear model which can be tested by means of some chi-squared test-statistics.

Special attention is given to the occurrence of *a priori* zero cell frequencies which have to be handled in a way quite different from the sampling zeros.

Contingency table analysis appears to be a useful tool for categorical data which are a result of regional economic survey data.

References

(1) Abrahamse, A.P.J. and W.M. Lammerts van Bueren, "Looking at Multiway Tables",Econometric Institute, Erasmus University (Report 7931), Rotterdam, 1979.

(2) Abrahamse, A.P.J. and W.M. Lammerts van Bueren, "Looking at Multiway Tables", Econometric Institute, Erasmus University (Report 8008), Rotterdam, 1980a.

(3) Abrahamse, A.P.J. and W.M. Lammerts van Bueren, "Nominal Variables: Dependence and Predictive Power", Econometric Institute, Erasmus University (Report 8017), Rotterdam, 1980b.

(4) Aitkin, M., "A Simultaneous Test Procedure for Contingency Table Models", Journal of the Royal Statistical Society, Series C, 28, no.3 (1979), pp. 233-242.

(5) Aitkin, M., "A Note on the Selection of Log-linear Models", Biometrics 36 (March 1980), pp. 173-178.

(6) Beggs, S. and S. Cardell, "Assessing the Potential Demand for Electric Cars", Journal of Econometrics 17, 1981, pp.1-21.

(7) Berkson, J., "Minimum Discrimination Information, the "No-Interaction" Problem and the Logistic Function", Biometrics 28, 1972, pp.443-468.

(8) Birch, M.W., "Maximum Likelihood in Three-Way Contingency Tables", Journal of the Royal Statistical Society 25, Series B, 1963, pp.220-233.

(9) Bishop, Y.M., S.E. Fienberg and P.W. Holland, Discrete Multivariate Analysis: Theory and Practice, M.I.T. Press, Cambridge, Massachusetts, 1975.

(10) Brouwer, F., "Log-linear Analyses with Metric and Non-metric Data: an Application to Spatial Data", Research Memorandum 1982-15, Department of Economics, Free University, Amsterdam, 1982.

(11) Brunden, M.N., "The Analysis of Non-independent 2x2 Tables from 2xc Tables using Rank Sums", Biometrics, 28, 1972, pp. 603-607.

(12) Clayton, D.G., "Some Odds Ratio Statistics for the Analysis of Ordered Categorical Data", Biometrika 61, no.3, 1974, pp. 525-531.

(13) Cox, D.R., The Analysis of Binary Data, Methuen, London, 1970.

(14) Everitt, B.S., The Analysis of Contingency Tables, Chapman and Hall, London, 1977.

(15) Fienberg, S.E., "An Iterative Procedure for Estimation in Contingency Tables", Annals Mathematical Statistics 41, 1970, pp. 907-917.

(16) Fienberg, S.E., The Analysis of Cross-Classified Categorical Data, M.I.T. Press, Cambridge, Massachusetts, 1977.

(17)Fienberg, S.E. and P.W. Holland, "Methods for Eliminating Zero Counts in Contingency Tables", in: Patil, G.P. (ed.). Random Counts in Scientific Work, The Pennsylvania State University Press, University Park, 1970.

(18) Fingleton, B., "Log-linear Modelling of Geographical Contingency Tables", Environment and Planning A 13, 1981, pp. 1539-1551.

(19)Gokhale, D. and S. Kullback, The Information in Contingency Tables, Marcel Dekker, New York, 1978.

(20)Goodman, L.A., "A Modified Multiple Regression Approach to the Analysis of Dichotomous Variables", American Sociological Review 37, 1972, pp. 28-46.

(21)Goodman, L.A., Analyzing Qualitative/Categorical Data, Abt Books, Cambridge, Massachusetts, 1978.

(22)Grizzle, J.E., C.F. Starmer and G.G. Koch, "Analysis of Categorical Data by Linear Models", Biometrics 25, 1969, pp.489-504.

(23)Grizzle, J.E. and O.D. Williams, "Log-linear Models and Tests of Independence for Contingency Tables", Biometrics 28, 1972, pp. 137-156.

(24)Israels, A.Z., "Multivariate Analyse Methoden voor Discrete Variabelen", Department of Statistical Methods, Central Office of Statistics, Voorburg, 1980.

(25)Israels, A.Z. and J. van Driel, "The Use of Chi-square Statistic for Selecting Explanatory Variables in Multiway Tables", Department of Statistical Methods, Central Bureau of Statistics, Voorburg, 1982.

(26)Kemper, F.J., "On Some Models of Categorical Data Analysis and their Application to Geographical Problems", Münstersche Geographische Arbeiten 11, (Bahrenberg, G. and U. Streit, ed.), Schöningh, Paderborn, 1981, pp. 45-58.

(27)Kullback, S., Information Theory and Statistics, Dover Publications, New York, 1968.

(28)McCullagh, P., "Regression Models for Ordinal Data", Journal of the Royal Statistical Society 42, no.2, 1980, pp. 109-142.

(29)Mood, A.M., F.A. Graybill and D.C. Boes, Introduction to the Theory of Statistics, McGraw Hill, London, 1974.

(30)O'Kelly, M.E., "Generalized Information Measures," Environment and Planning A 13, 1981, pp. 681-688.

(31)Pannekoek, J., "A Linear and a Log-linear Model for the Analysis of Ratingscale Data", Department of Statistical Methods, Central Bureau of Statistics, Voorburg, 1980.

(32)Pannekoek, J.,"Informatiematen, Associatie en Log-lineaire Modellen", Department of Statistical Methods, Central Bureau of Statistics, Voorburg, 1981.

(33)Pannekoek, J. and L.H. Stronkhorst, "Exploraties met behulp van log-lineaire Modellen", Department of Statistical Methods, Central Bureau of Statistics, Voorburg, 1981.

(34)Patel, J.K., C.H. Kapadia and D.B. Owen, Handbook of Statistical Distributions, Marcel Dekker, New York, 1976.

(35)Payne, C., "The Log-linear Model for Contingency Tables", in: C. Payne and C.O. Muircheartaigh (eds), The Analysis of Survey Data, vol.2: Model Fitting, Wiley, New York, 1977, pp.105-144.

(36)Placket, R.L., The Analysis of Categorical Data, Griffin, London, 1974.

(37)Scholten, H., "Describing and Forecasting Migration within a Housing Market Area: the Log-linear Model used as a Spatial Interaction Model, paper presented at the Third European Colloquium on "Theoretical and Quantitave Geography", Augsburg, 1982.

(38)Theil, H.,"On the Estimation of Relationships involving Qualitative Variables", American Journal of Sociology 76, 1970, pp. 103-154.

(39)Willekens, F., "Entropy, Multiproportional Adjustement and Analysis of Contingency Tables", Sistemi Urbani 2/3, 1980, pp. 171-201.

(40)Willekens, F.,"Log-linear Modelling of Spatial Interaction", paper presented at the 22-nd European Congress of the Regional Science Association, Groningen, 1982.

(41)Willekens, F., A.Pór and R. Raquillet, "Entropy, Multiproportional and Quadratic Techniques for inferring Patterns of Migration from Aggregate Data", IIASA Reports 4, no.1, Laxenburg, 1981, pp.83-124.

(42)Wrigley, N., "Developments in the Statistical Analysis of Categorical Data, "Progress in Human Geography 3, 1979, pp. 315-355.

(43)Wrigley, N., "Categorical Data, Repeated-Measurement Research Designs and Regional Industrial Surveys", Regional Studies 14, 1980, pp. 455-471.

(44)Wrigley, N., "Quantitative Methods: a View on the Wider Scene", Progress in Human Geography 5, no.4, 1981, pp. 548-561.

Serie Research Memoranda

1980-1 jan.	P. Nijkamp and H. Voogd	New Multicriteria Methods for Physical Planning by Means of Multidimensional Scaling Techniques
1980-2	Hidde P. Smit	Medium- and Long-Term Models for the Escap-Region - A Review of existing models and a proposal for a new model system -
1980-3 april	P.v. Dijck en H.Verbruggen	Productive Employment in Developing Countries' Exporting Industries
1980-4	P. Nijkamp and L. Hordijk	Integrated Approaches to Regional Development Models; A survey of some Western European Models
1980-5	P. Nijkamp	Soft Econometric Models; An Analysis of Regional Income Determinants
1980-6	P. Nijkamp and F.van Dijk	Analysis of Conflicts in Dynamical Environmental Systems via Catastrophe Theory
1980-7 juni	E. Vogelvang	A short term econometric model for the Consumer demand of roasted Coffee in The Netherlands
1980-8	N.van Hulst	De effectiviteit van de geleide loonpolitiek in Nederland
1980-9 okt.	P. Nijkamp	A survey of Dutch integrated Energy-Environmental-Economic Policy Models
1980-10	P. Nijkamp	Perspectives for Urban Analyses and Policies
1980-11	P. Nijkamp	New developments in Multidimensional geographical data and Policy analysis
1980-12	F.C. Palm, E. Vogelvang and D.A. Kodde	Efficient Estimation of the Geometric Distributed Lag Model; some Monte Carlo Results on Small Sample Properties
1981-1 jan.	E. Vogelvang	A quarterly econometric model for the Price Formation of Coffee on the world market
1981-2	Hidde P. Smit	Demand and Supply of natural Rubber, Part I
1981-3	R. Vos	The political Economy of the Republic of Korea; A proposal for a model framework of an open economy in the Escap-region, with emphasis on the role of the State
1981-4	F.C. Palm	Structural Econometric Modeling and Time Series Analysis - Towards an Integrated Approach
1981-5	P. Nijkamp in co-op. with H.van Handenhoven and R. Janssen	Urban Impact Analysis in a Spatial Context: Methodology and Case Study
1981.6	R. Ruben	Primaire exporten en ekonomiese ontwikkeling
1981-7 maart	D.A. Kodde	Het genereren en evalueren van voorspellingen van omzet en netto winst: een toegepast kwantitatieve benadering
1981-8	B. Out	Financiële vraagstukken onder onzekerheid
1981-9	P.van Dijck and H.Verbruggen	A Constant-Market-Shares Analysis of ASEAN Manufactured Exports to the European Community